PAGE No.
DATE / /

Assignment Building and Analyzing Transformer-Based Model's.

1. Self-attention allows a model to relate different positions of a single sequence to compute a representation of that same sequence. It helps the model understand which words are most relevant to other's, regardless of their distance.

Example Sentence "The animal didn't cross the street because it was too tired".
Here, self-attention allows the model to associate 'it' with 'animal' rather than 'street'

The Mechanism (Q, K, V)
For every input word embedding, three vectors are created.
1. Query (Q): What I am looking for.
2. Key (K): What I contain (used for matching)
3. Value (V): The information I actually provide

The attention weight determines how much focus to place on other part's of the sentence
It is calculated using the Scaled Dot-Product Attention formula.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

PAGE No.
DATE / /

2 Model Type Encoder-only (e.g. BERT)
Architecture Bidirectional attention
Best for Understanding, classification
Mechanism Can see words to the left and right simultaneously

Model-Type Decoder-only (e.g. GPT)
Architecture Unidirectional (causal) attention.
Best for Generative tasks, Text completion.
Mechanism Can only see previous words.

Model-Type Encoder-Decoder (eg. T5)
Architecture Hybrid; Encoder processes input, decoder generates output.
Best For Translation, Summarization.
Mechanism Encoder "thinks", then passes context to the Decoder.

3. This paradigm works because of Transfer Learning a model learns general linguistic patterns on a massive dataset before being sepecialized for a specific task.
- Masked Language Modeling (MLM) Used in pre-training (eg. BERT). Words are hidden (masked), and the model learns to predict them using surrounding context. This build's deep bidirectional understanding
- Causal Language Modeling (CLM) Used in generative models (e.g GPT). The model predicts the next word in a sequence.

PAGE No.
DATE / /

This teaches the model the "flow" and logic of language generation.

- The Benefit: Pre-training acts as a "head start". Insted of learning English from scratch for a small task like sentiment analysis, the model already understands grammer, facts, and nuances. Fine-tuning then just adjusts the final layers to map the vast knowledge to a specific output.